



THE UNIVERSITY  
*of*ADELAIDE



BIOINFORMATICS AND SYSTEMS MODELLING

## Ancient DNA

Bastien Llamas

BIOINF 3000 / BIOTECH 7005



@DNATimeTravel



bastien.llamas@adelaide.edu.au



- Introduction to ancient DNA
- Sources of ancient DNA and post-mortem DNA decay
- Properties of ancient DNA
- Ancient DNA analysis

## Lecture overview

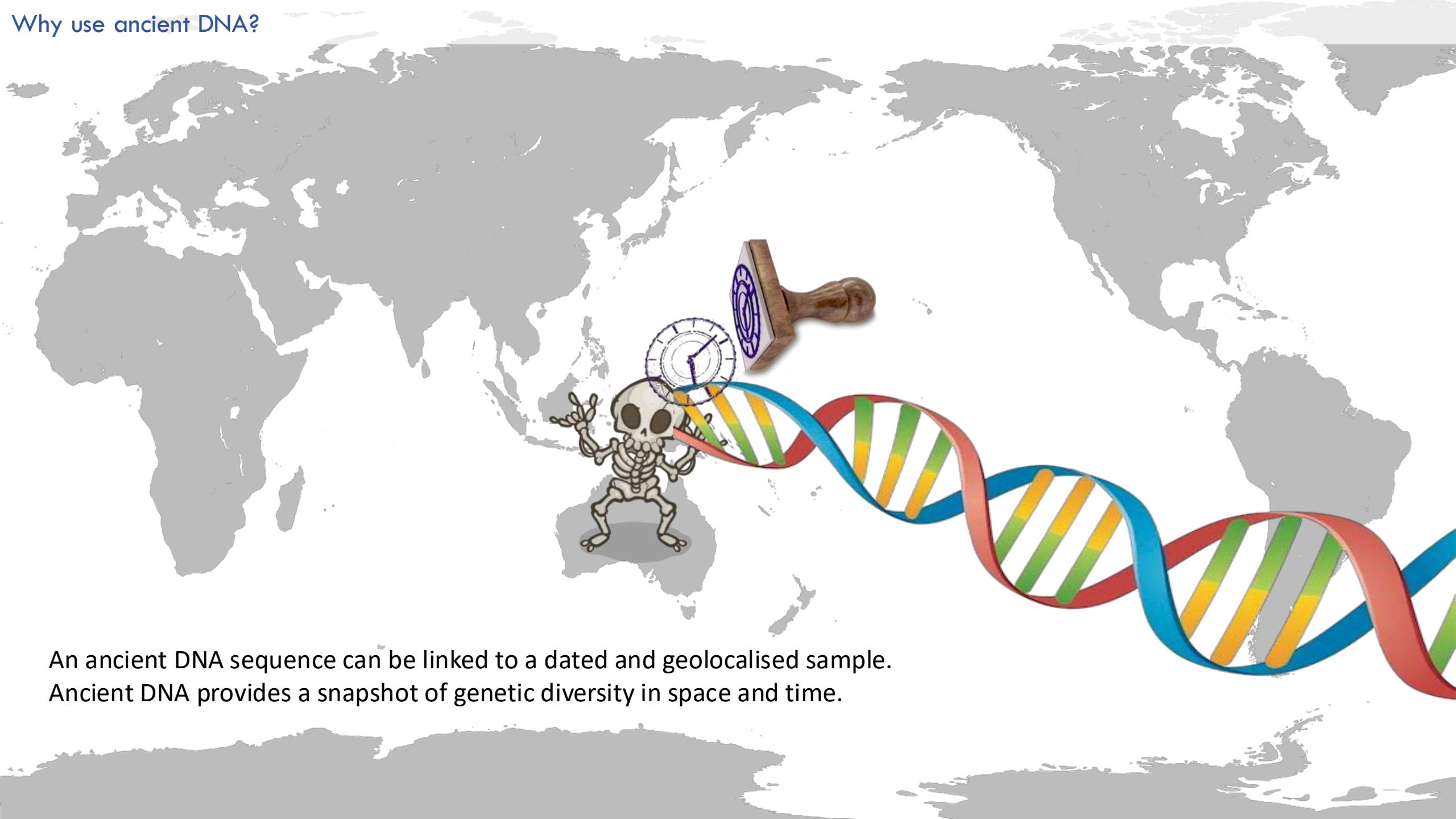
- Introduction to ancient DNA
- Sources of ancient DNA and post-mortem DNA decay
- Properties of ancient DNA
- Ancient DNA analysis

## Ancient DNA

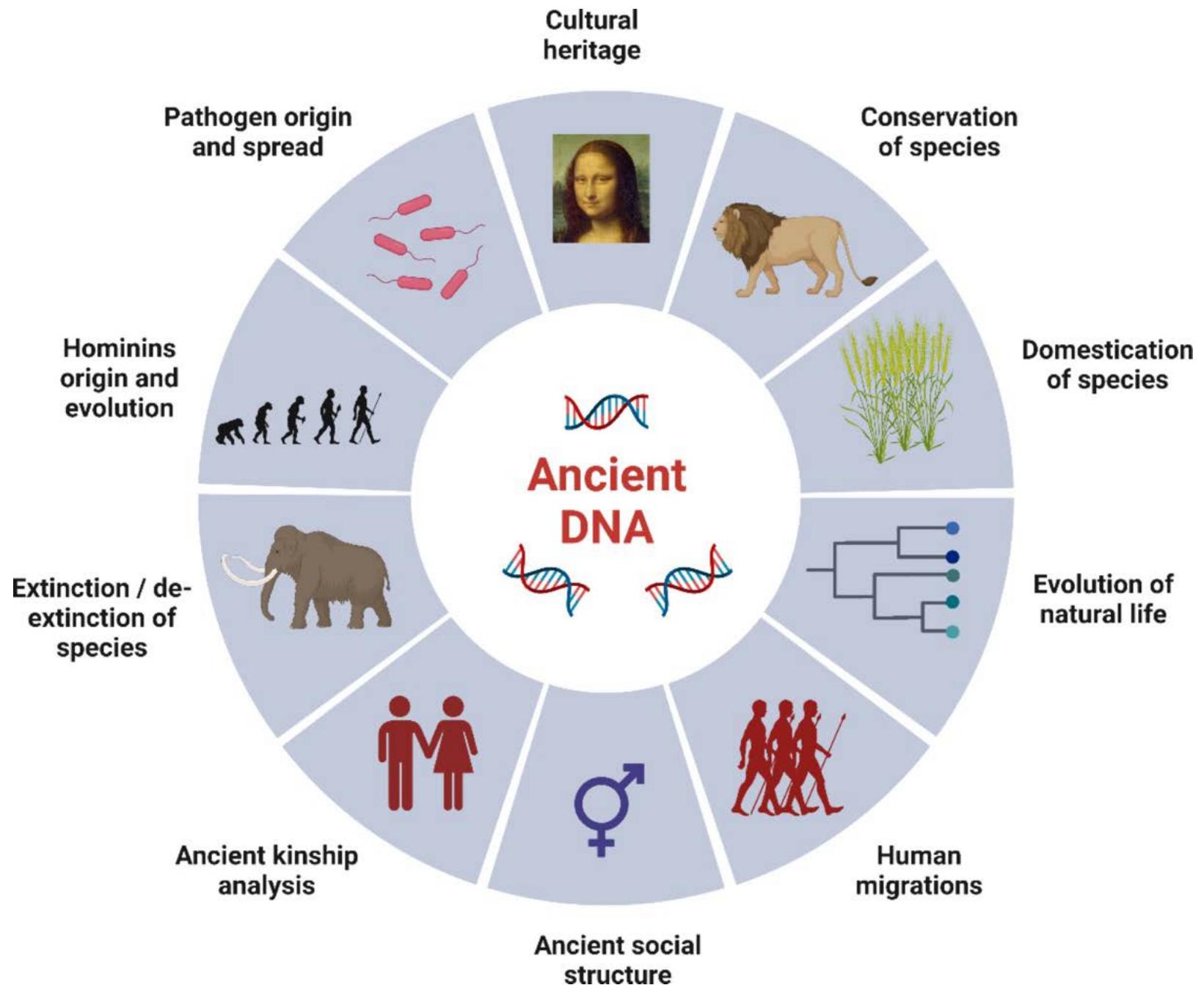
- Ancient DNA is extracted from archaeological (or palaeontological) organic material
- Ancient DNA can survive thousands of years but is challenging to study



## Why use ancient DNA?

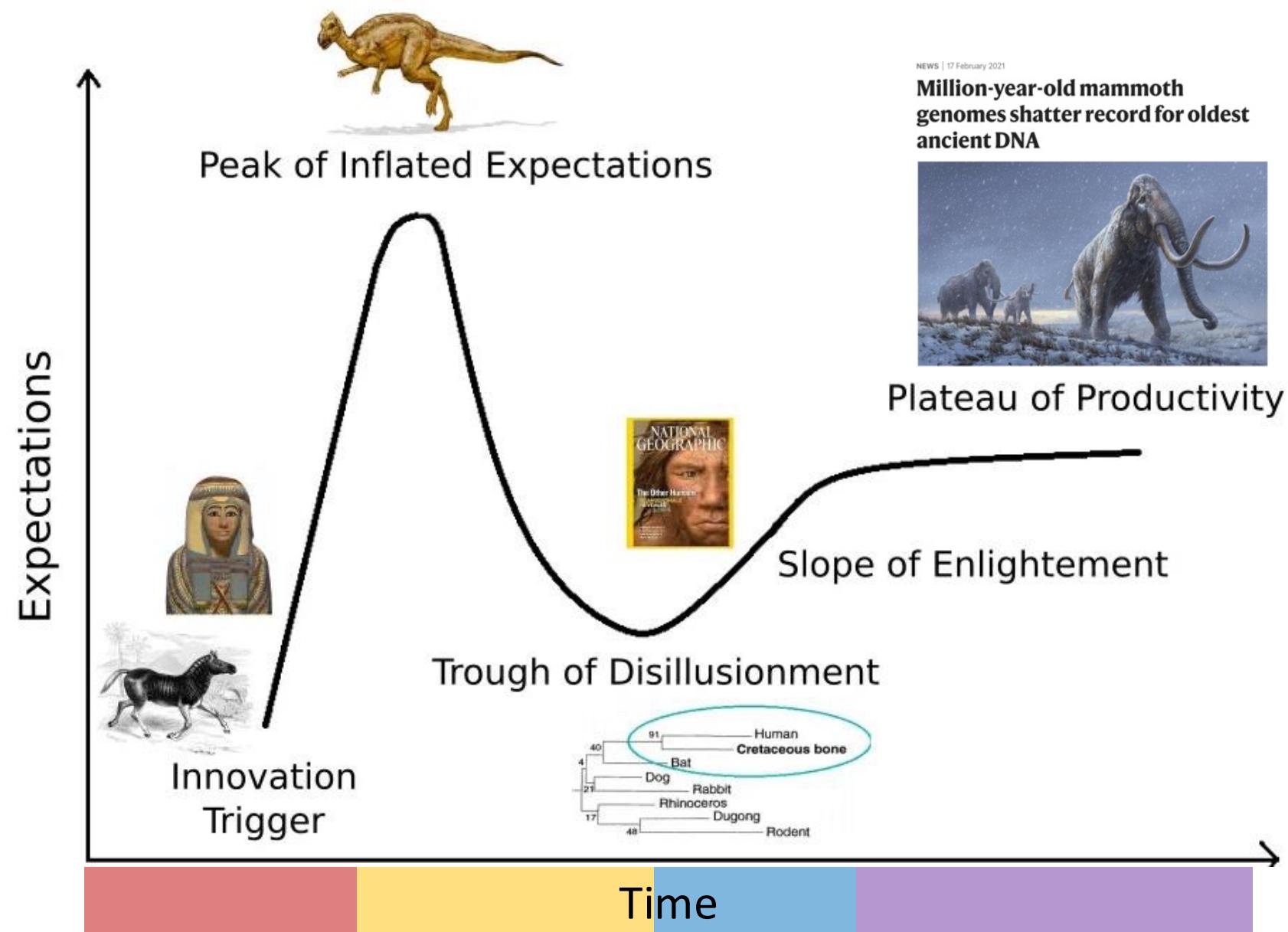


An ancient DNA sequence can be linked to a dated and geolocalised sample.  
Ancient DNA provides a snapshot of genetic diversity in space and time.



# Ancient DNA is coming of age

- First ancient DNA studies:
  - ✓ Higuchi et al. (1984) Nature
  - ✓ Pääbo et al. (1985) Nature
- Discredited study claiming dinosaur ancient DNA:
  - ✓ Woodward et al. (1994) Science
- First draft of the Neanderthal genome:
  - ✓ Green et al. (2010) Science
- First genome over a million years old:
  - ✓ Van der Valk et al. (2021) Nature



NEWS | 17 February 2021

**Million-year-old mammoth genomes shatter record for oldest ancient DNA**



Plateau of Productivity

- Introduction to ancient DNA
- Sources of ancient DNA and post-mortem DNA decay
- Properties of ancient DNA
- Ancient DNA analysis

# Sources of ancient DNA

BONES



DENTAL CALCULUS



MAIZE COBS



MOLLUSC SHELLS



TEETH



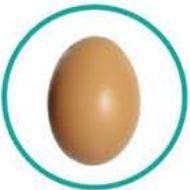
MUSEUM SKINS



GRAPE VINE & SEEDS



EGGSHELLS



BULK-BONE



COPROLITES



HERBARIUM



PARCHMENT



MUMMIES



POLLEN



HAIRS



CLOTHING



SEDIMENT CORES

# Sampling is intrinsically linked to archaeology



Supe Valley, Peru

Credit: Bastien Llamas



Credit: Linda Manzanilla



Credit: Bastien Llamas



## Post-mortem DNA decay and contamination

Past

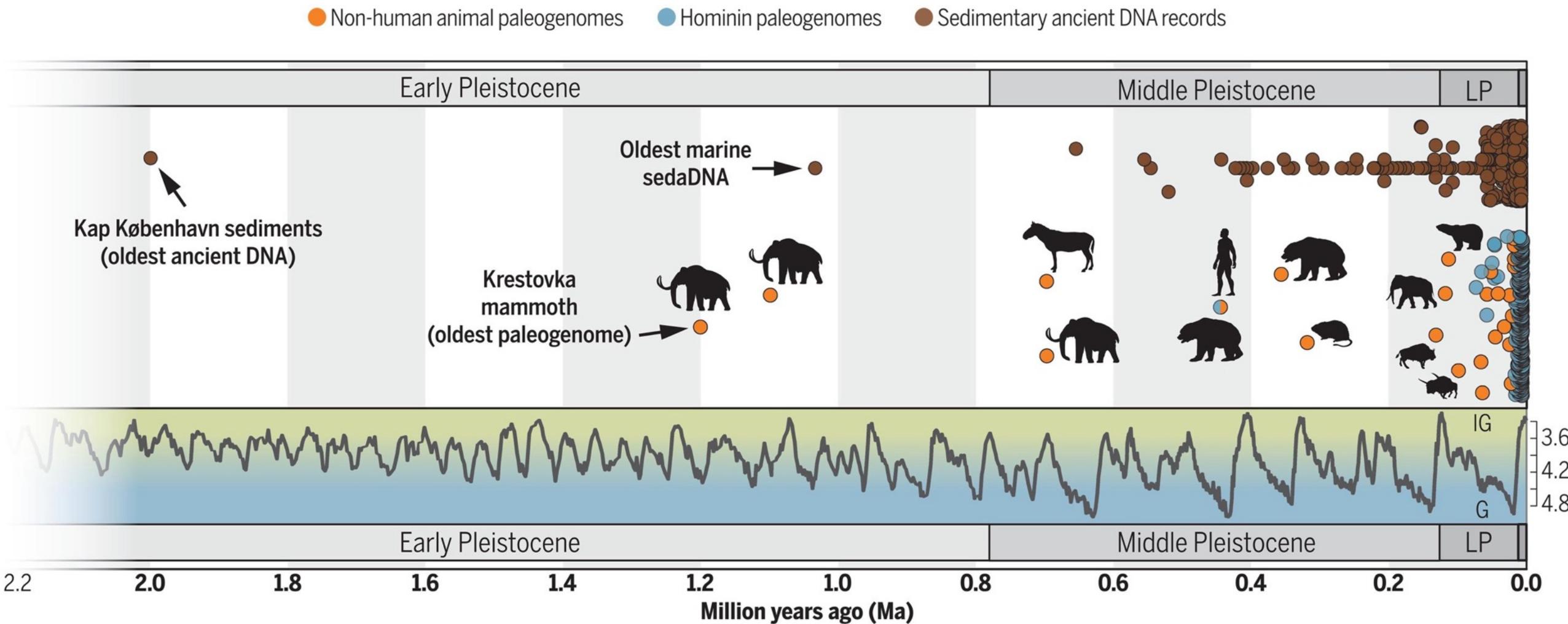


Present



- DNA decay starts immediately after death due to endonucleases and microorganisms
- Over time, hydrolytic and oxidative processes will keep degrading DNA
- Favourable conditions (low temperature, anoxic, dry) help preserve DNA over thousands of years

# Empirical time limit of DNA preservation

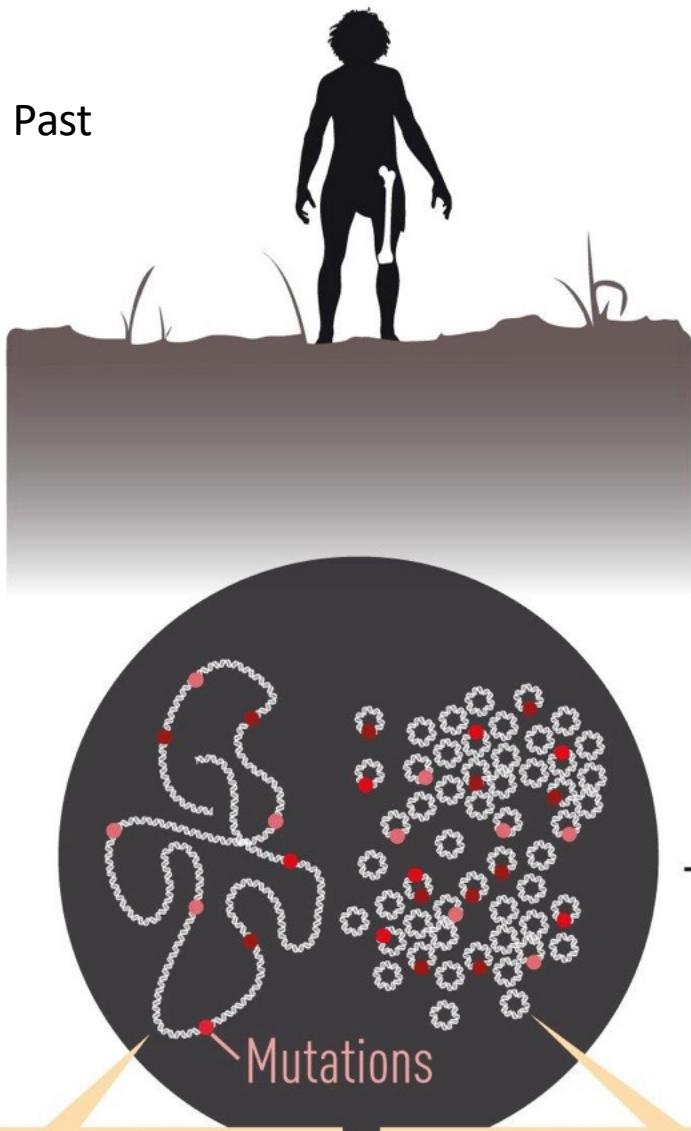


Empirical time limit of DNA preservation is ~1.5 million years



- Introduction to ancient DNA
- Sources of ancient DNA and post-mortem DNA decay
- Properties of ancient DNA
- Ancient DNA analysis

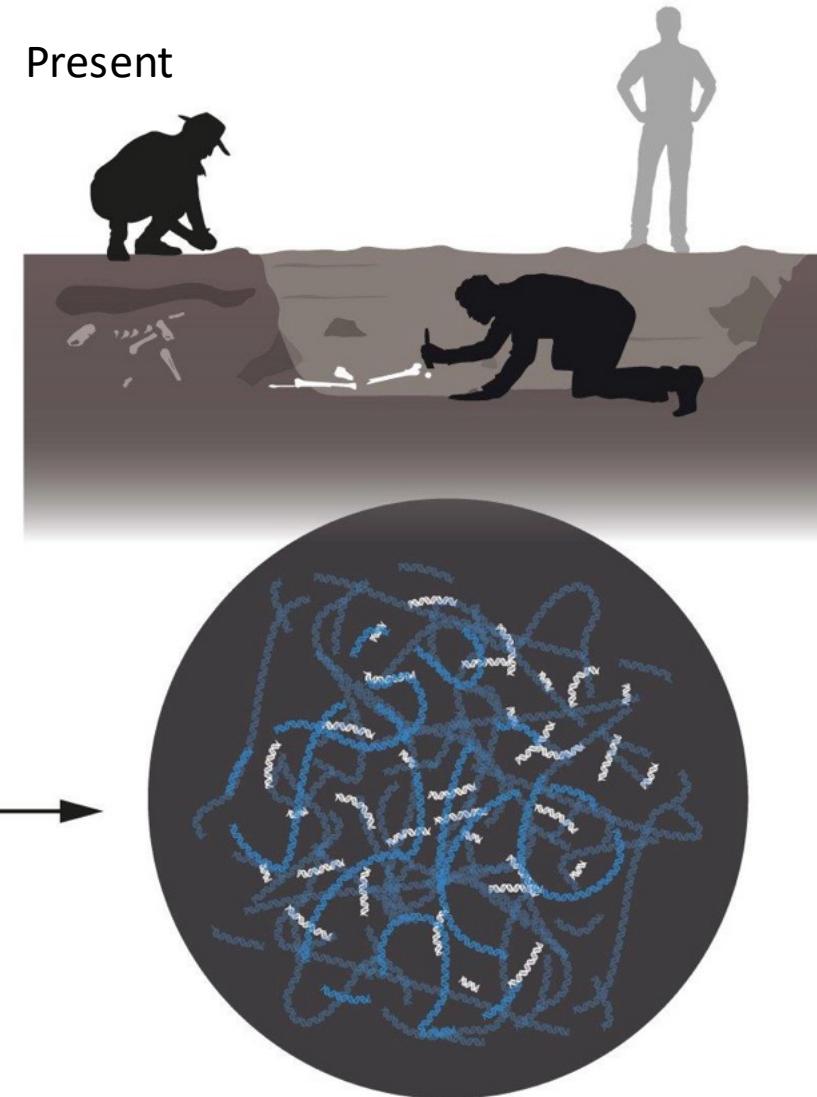
## The three main properties of ancient DNA



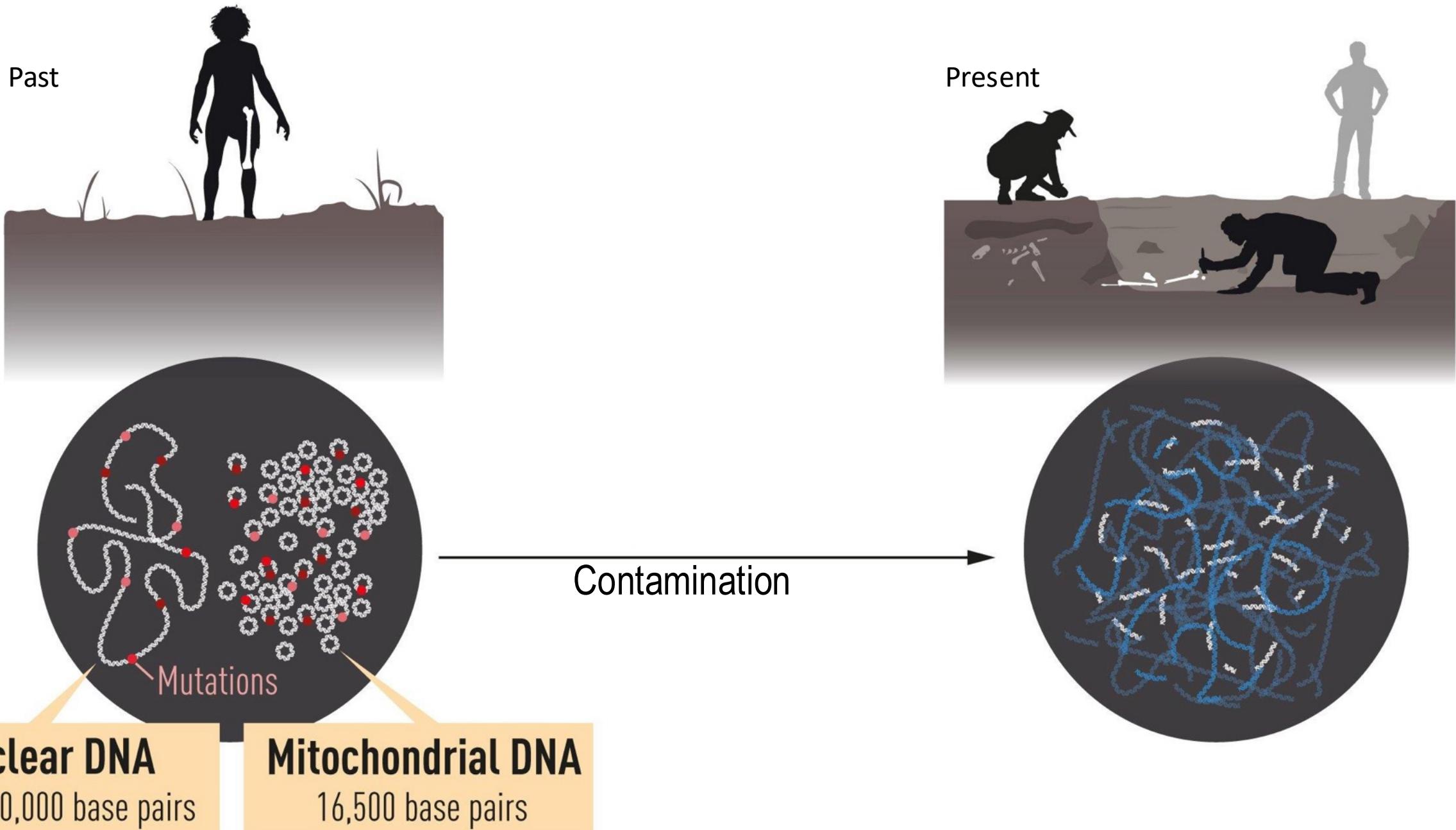
**Nuclear DNA**  
3,000,000,000 base pairs

**Mitochondrial DNA**  
16,500 base pairs

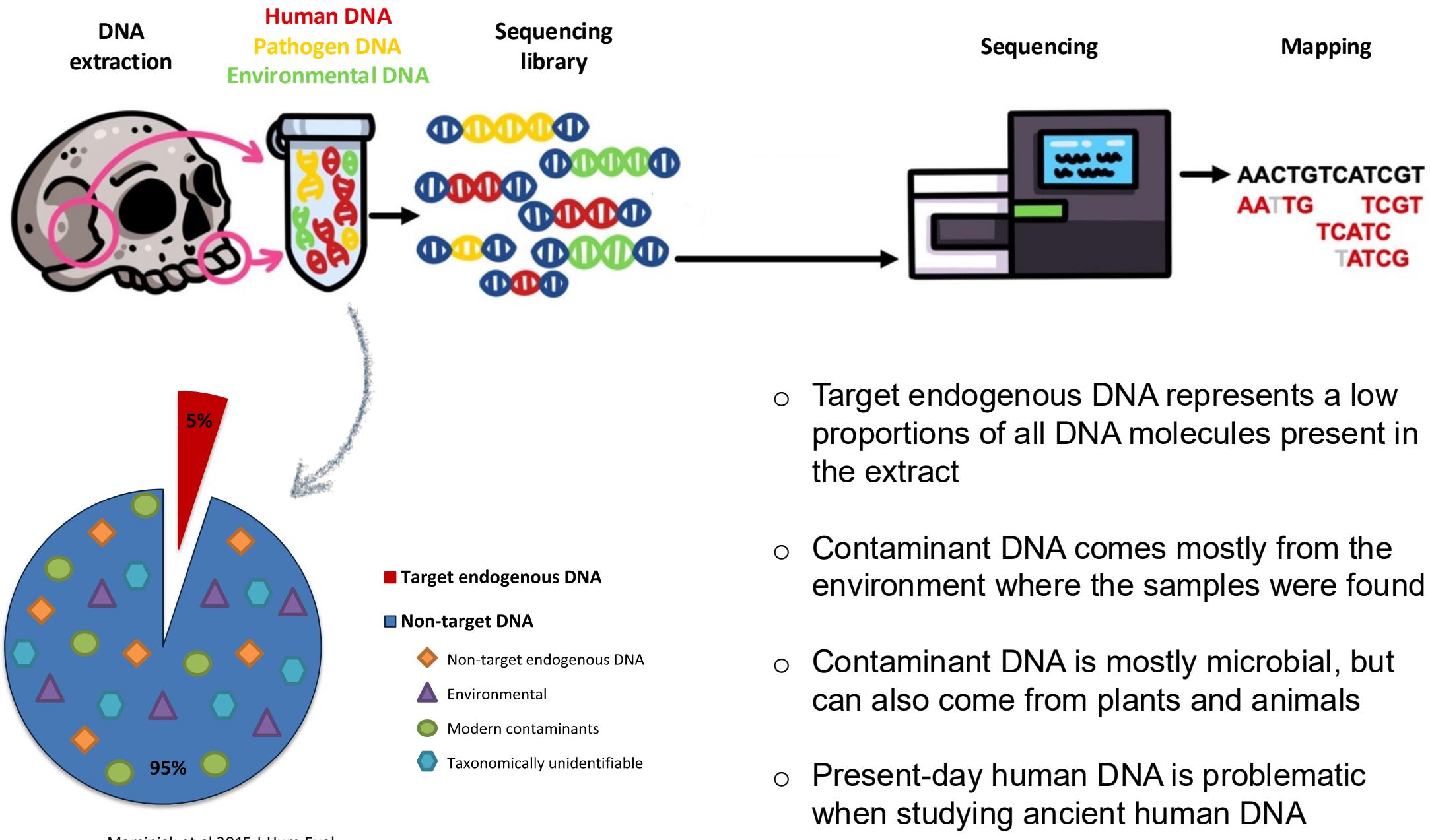
Contamination  
Base modifications  
Fragmentation



## The first main property of ancient DNA: contamination



# The issue of DNA contaminants from the environment



## Protect the sample from yourself



Credit: Guido Valverde



Credit: Bastien Llamas



Credit: Linda Manzanilla



Credit: Raphael Eisenhofer



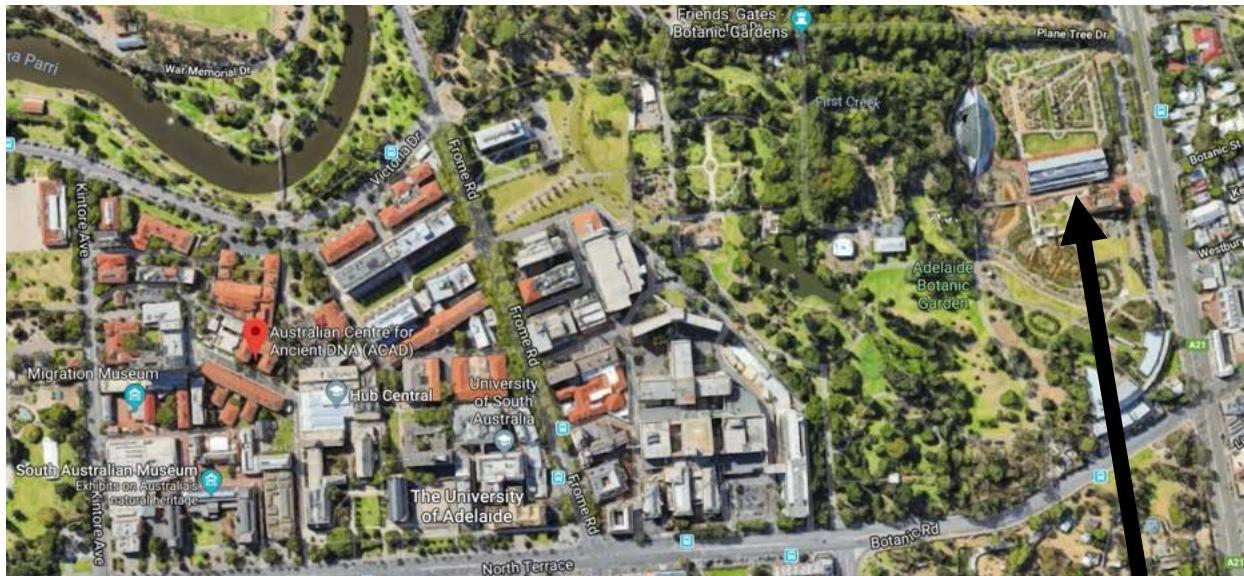
Credit: Julien Soubrier



Credit: Julien Soubrier

PPE in the field and in the lab

# The Australian Centre for Ancient DNA: a low-DNA environment



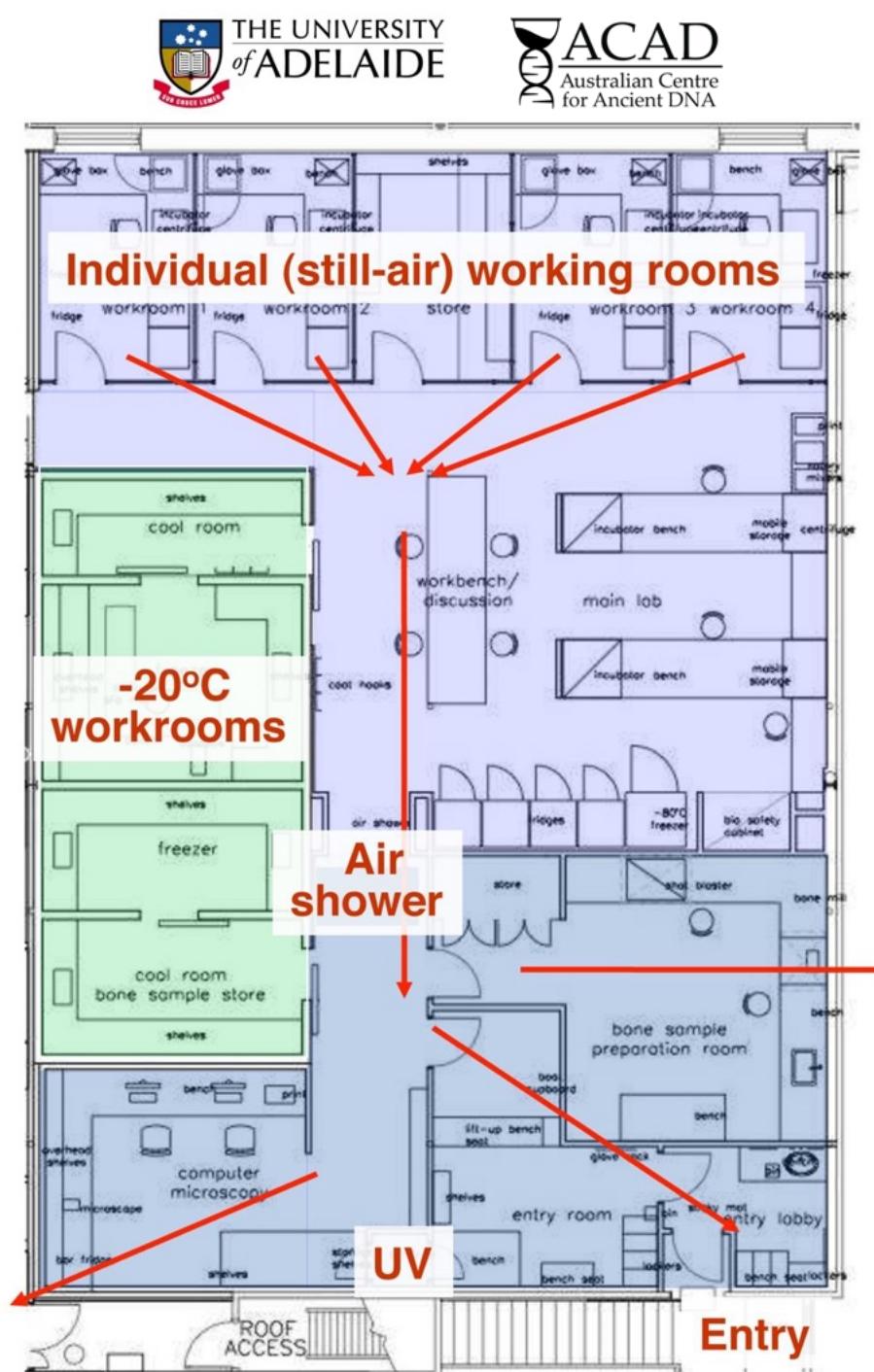
Google Maps



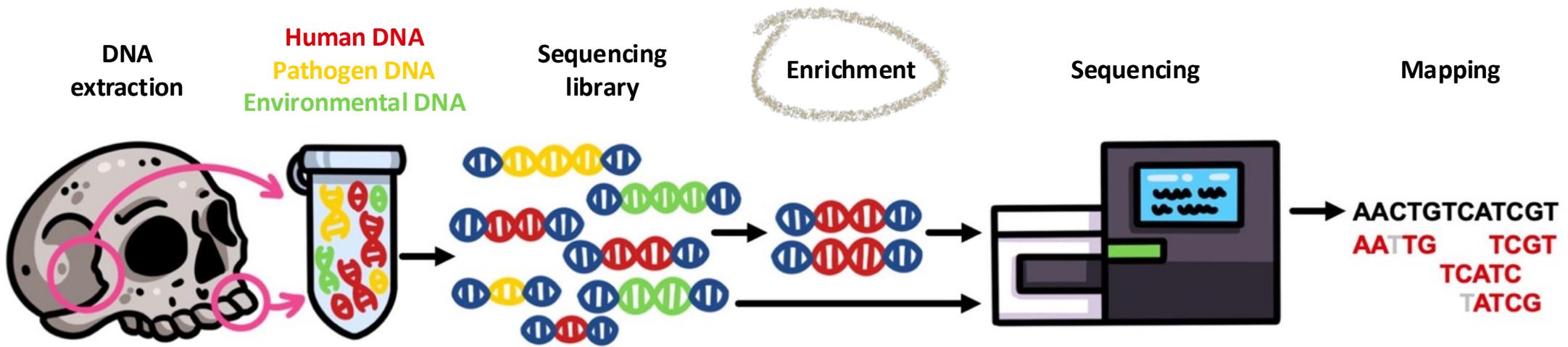
Credit: Jeremy Austin

## Specialist laboratory facilities

- Clean rooms physically isolated from sources of DNA contaminants:
    - No amplification of DNA
    - Positive air pressure
  - Clean rooms dedicated to specific tasks:
    - Sample storage
    - Sample preparation
    - DNA extraction
    - Molecular biology (pre-amplification)
  - UV and bleach treatment of work surfaces
  - DNA-free reagents and consumables
  - PPE and strict protocols



## Enrichment of target DNA

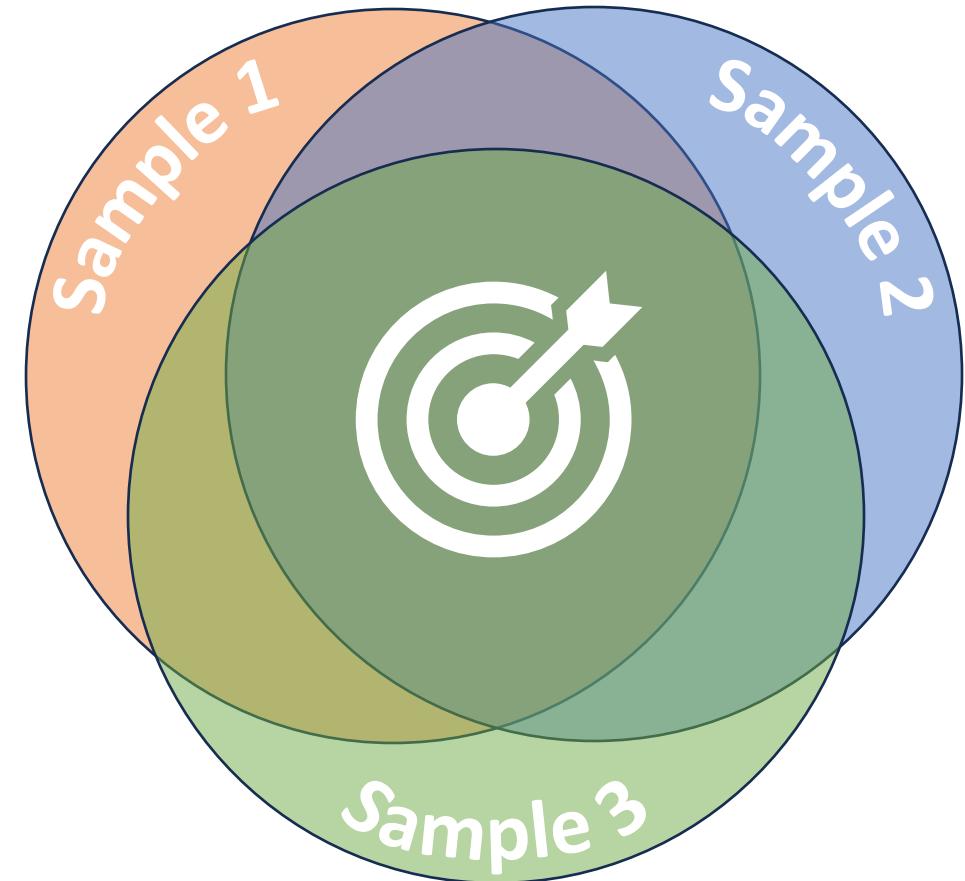


- Enrich target endogenous DNA using hybridisation to DNA or RNA baits
- Target endogenous DNA can be:
  - The whole genome
  - A chromosome
  - The mitochondrial genome
  - All the genes (the exome)
  - Genome-wide variable positions (Single Nucleotide polymorphisms—SNPs)
  - chrY SNPs

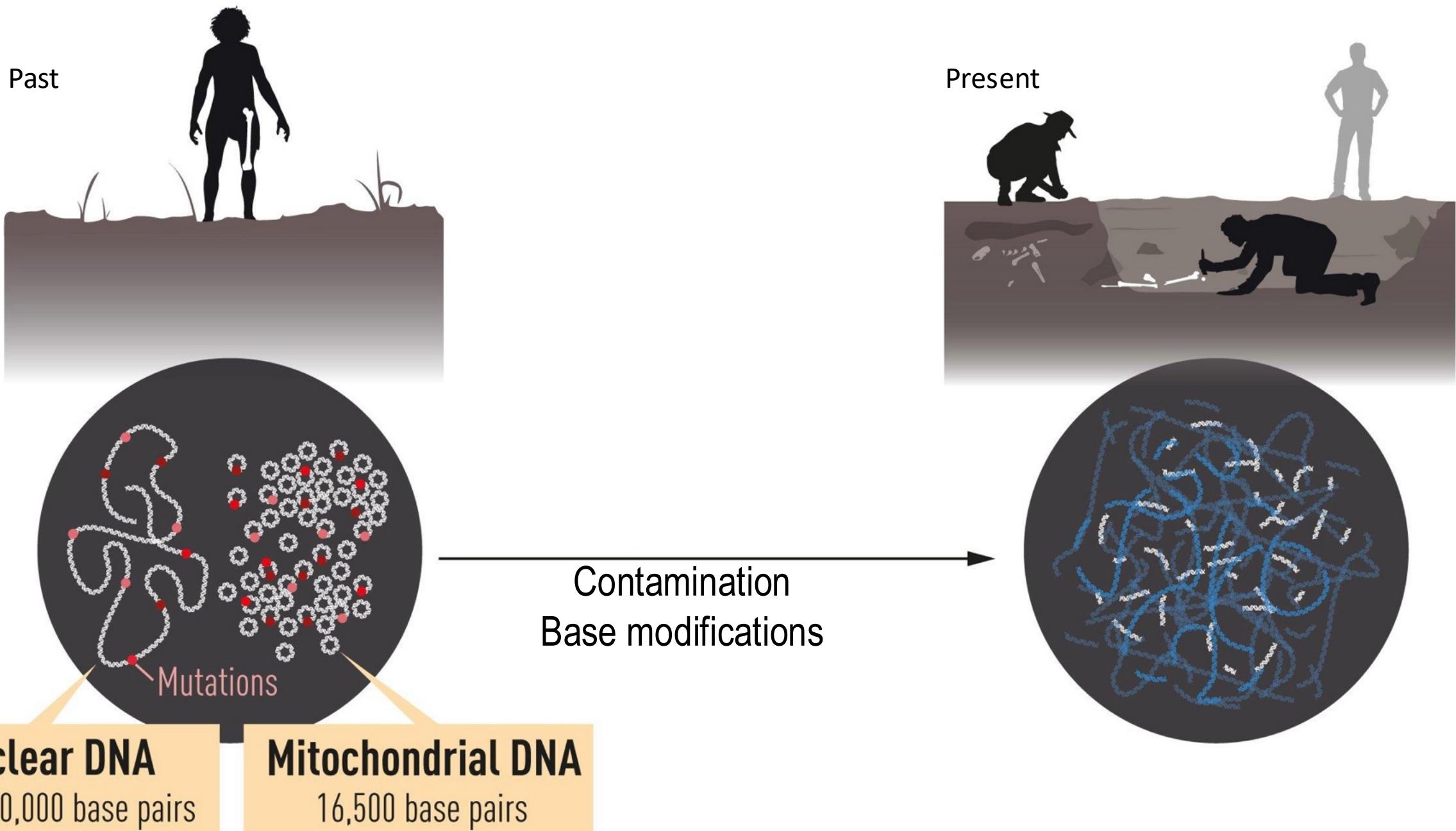
## Benefits of target DNA enrichment

Benefits include:

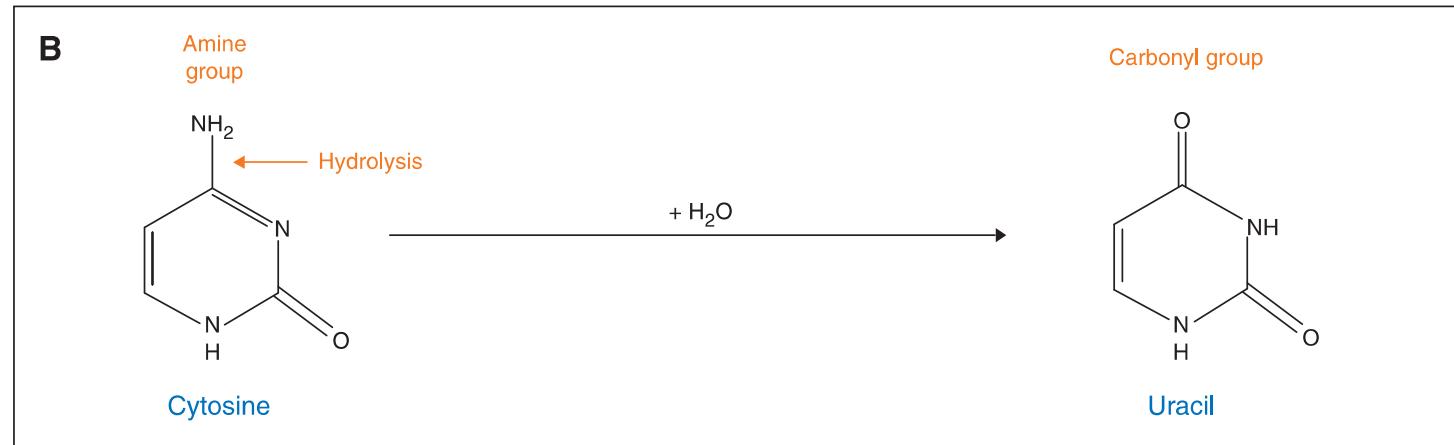
- Increase the proportion of endogenous DNA in the sequencing data
- Increase the proportion of endogenous DNA common between samples
- Decreased sequencing costs



## The second main property of ancient DNA: base modifications



## Ancient DNA damage: cytosine deamination

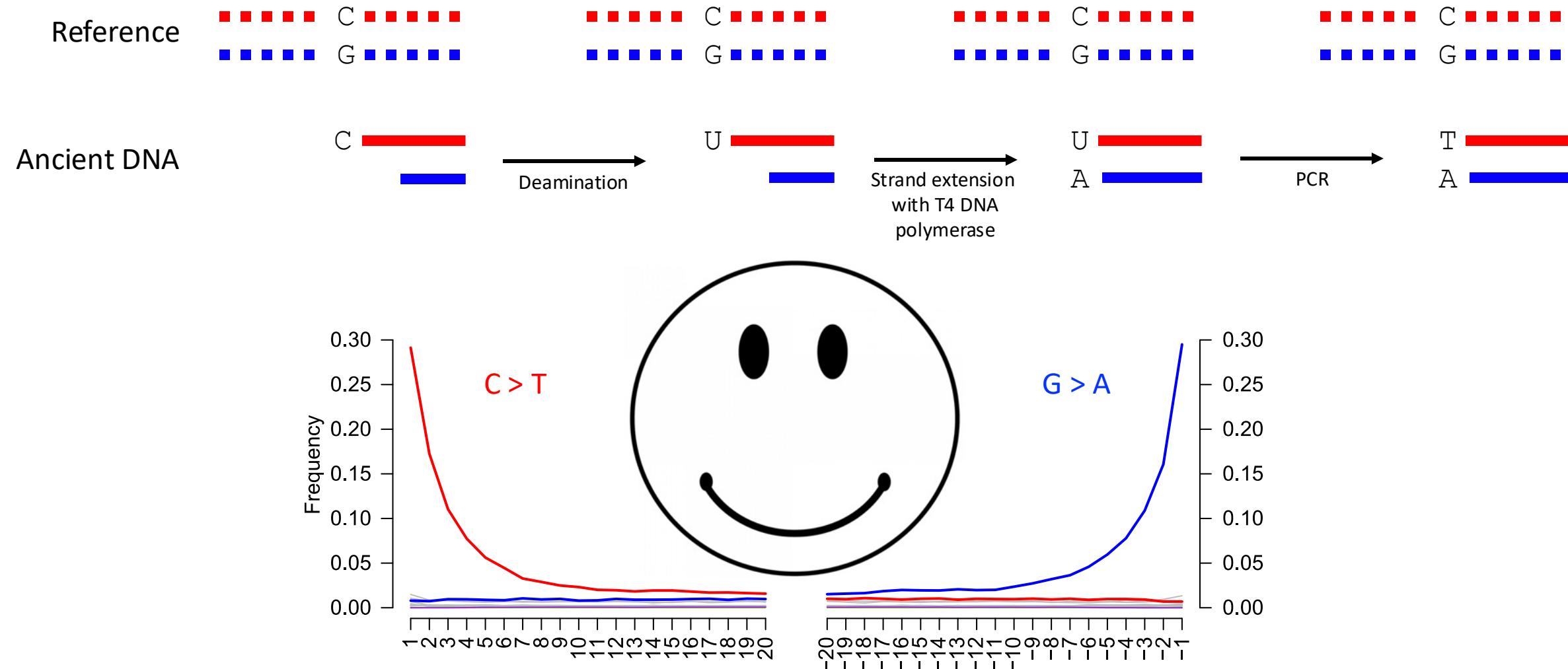


Dabney et al 2013 Cold Spring Harbor perspectives in biology

- Deamination of cytosine to uracil is the major mechanism leading to miscoding lesions in ancient DNA
  - If not dealt with experimentally or bioinformatically, miscoding lesions can lead to false positive results

## Cytosine deamination accumulates at ancient molecules ends

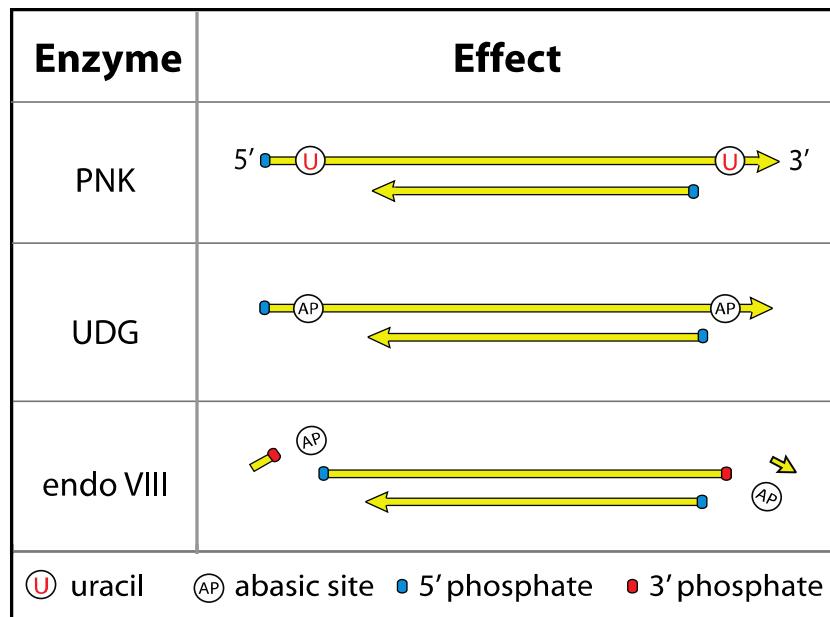
DNA polymerases will incorporate an A across from the U, and in turn a T across from the A, causing apparent G to A and C to T substitutions



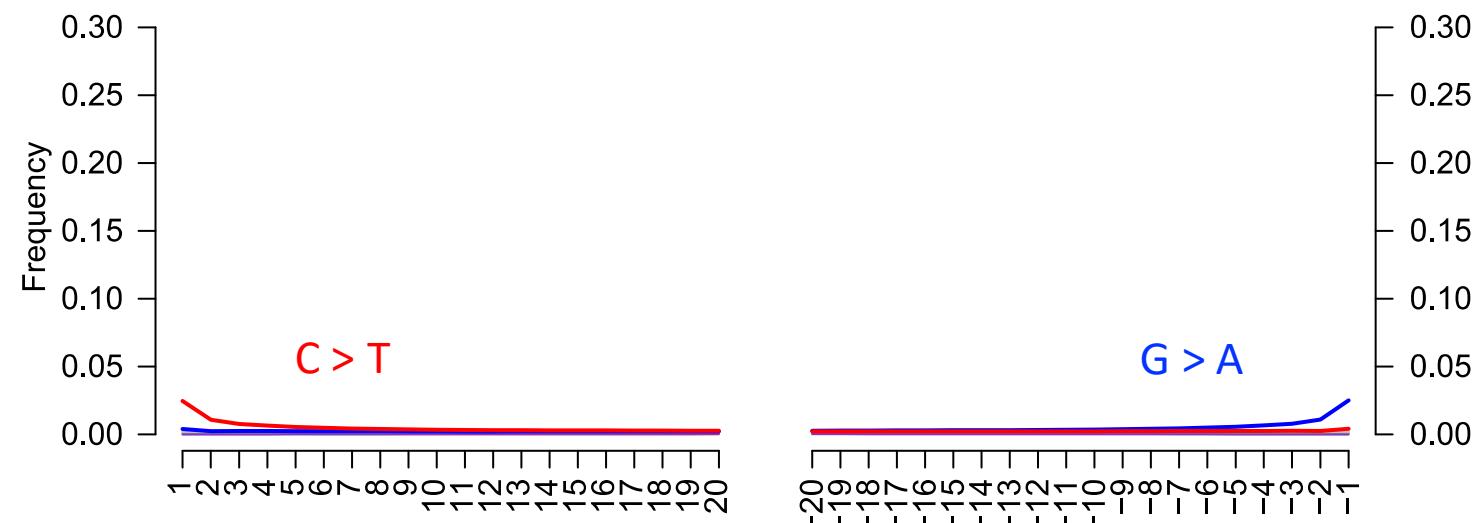
## Uracils can be removed experimentally

DNA can be “repaired” before preparing sequencing libraries by:

- Removing the uracil base with UDG (Uracil-DNA Glycosylase), leaving an abasic site
- Cleaving the DNA backbone at the 3' and 5' sides of the abasic site with endonuclease VIII



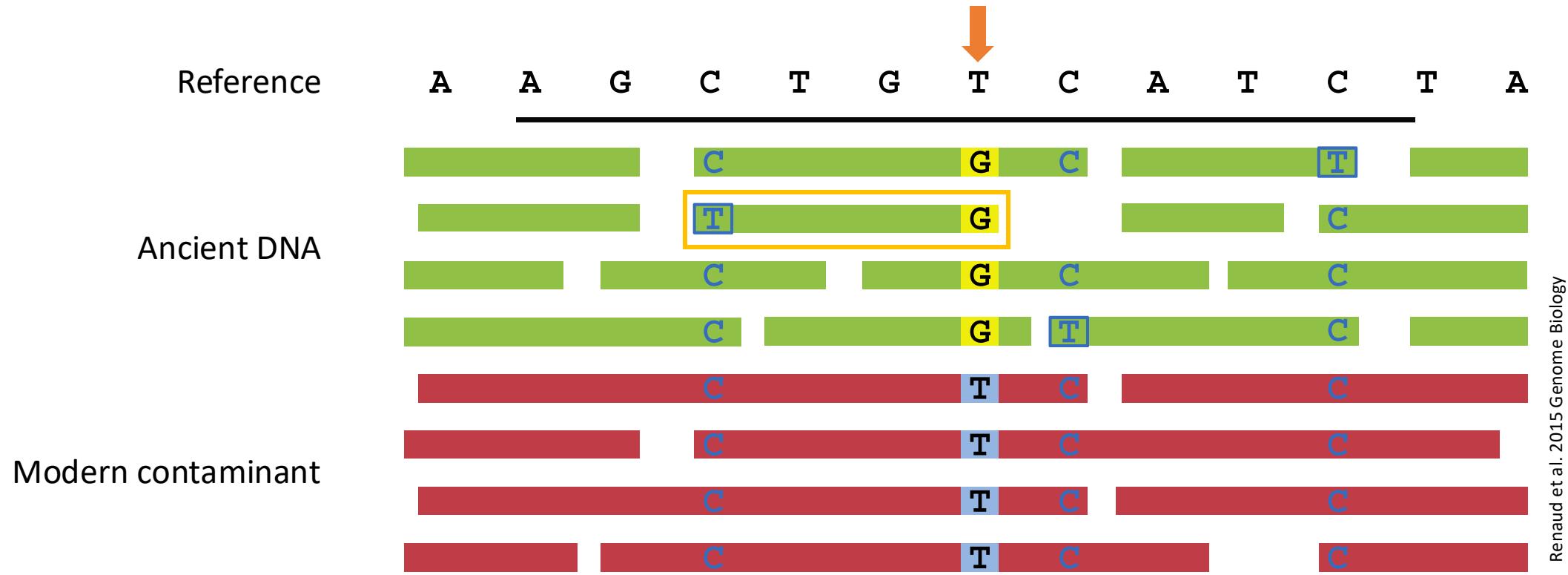
Adapted from Briggs et al. 2010 Nucleic Acids Research



Consequences of DNA repair:

- No accumulation of C-to-T and G-to-A substitutions at the DNA fragment ends
- Shorter DNA fragments

## Contamination estimation from haploid DNA

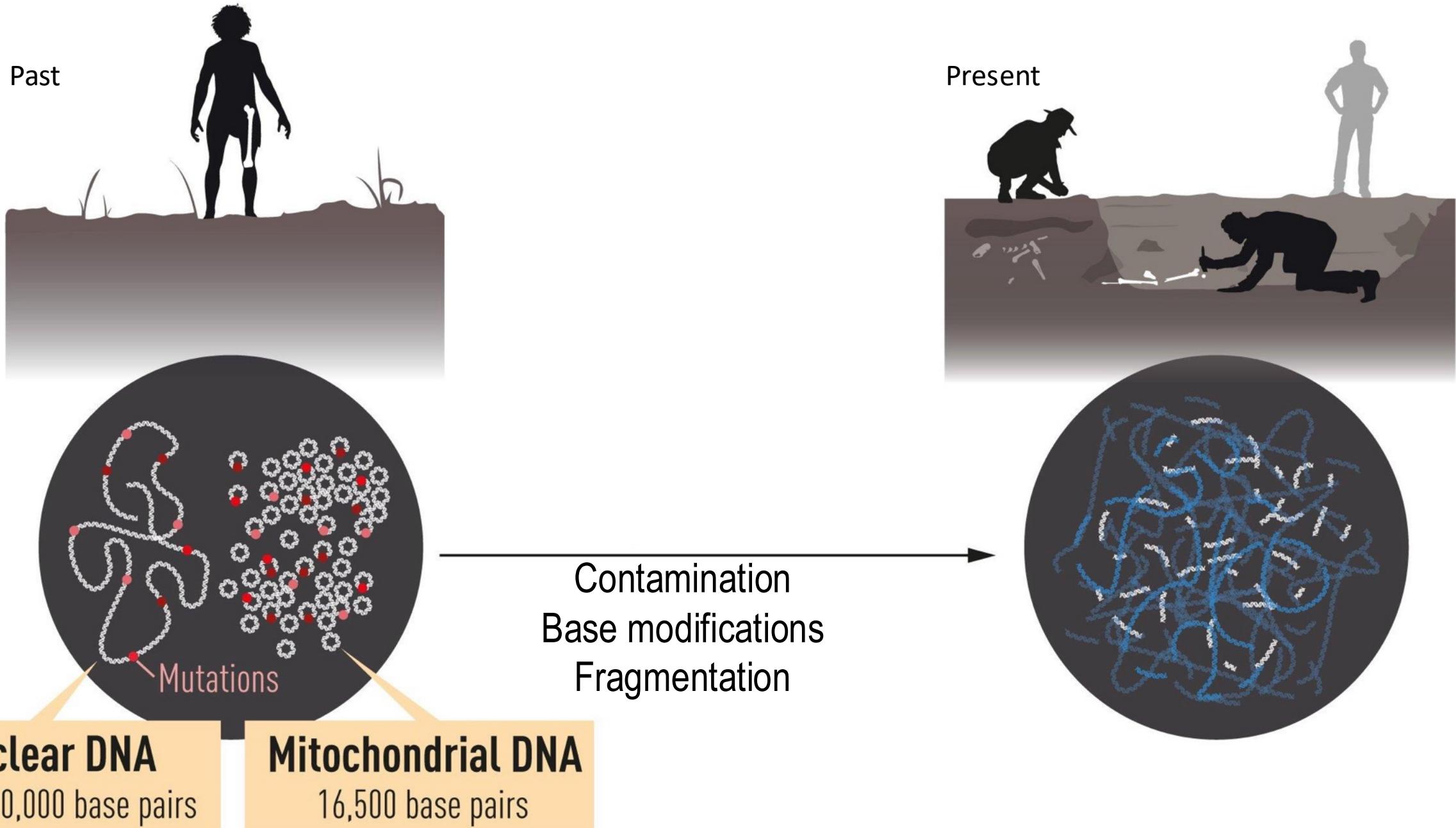


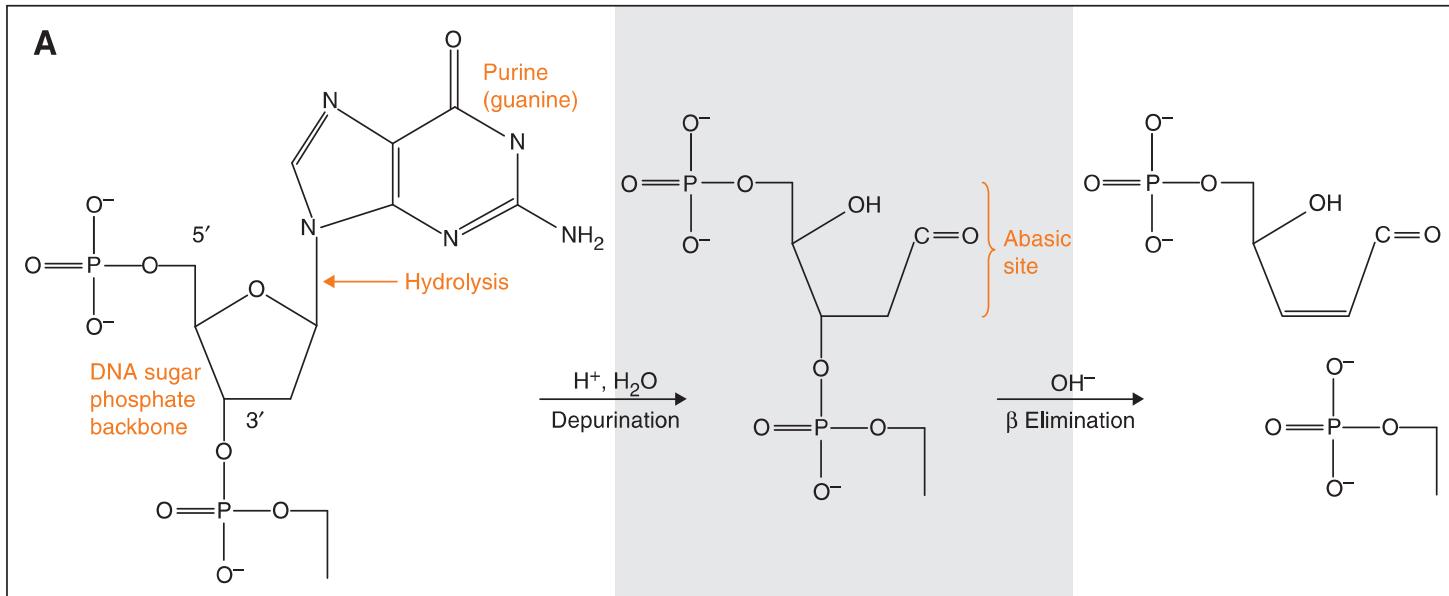
Renaud et al. 2015 Genome Biology

In data mapped to haploid genomes (mitochondrial DNA in both sexes or X chromosome in males):

1. Identify heteroplasmic positions (here G/T, orange down arrow)
2. Identify ancient DNA fragments that carry deaminated cytosines near the ends of the DNA fragments (T in blue squares)
3. The ancient allele is linked to cytosine deamination in some ancient DNA fragments (fragment in the orange box). Estimate contamination from fragments that do not contain the ancient allele

## The third main property of ancient DNA: fragmentation



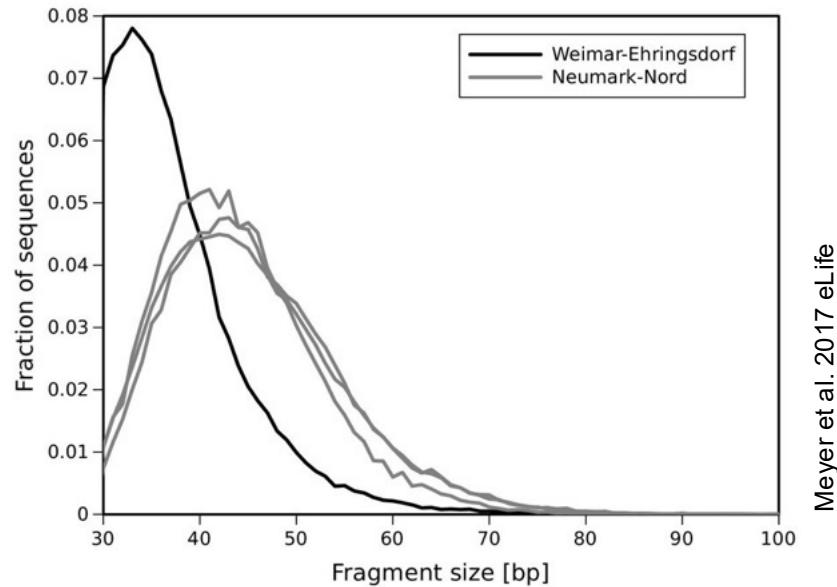


Dabney et al 2013 Cold Spring Harbor perspectives in biology

## Steps to DNA fragmentation:

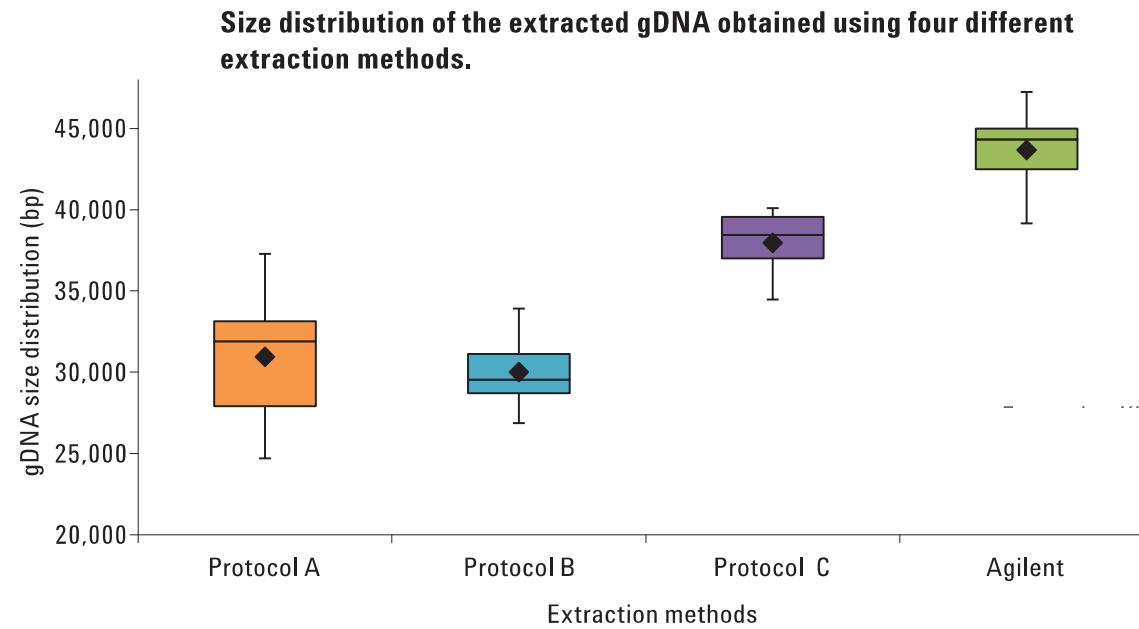
- Hydrolytic depurination (N-glycosyl bond between sugar and adenine or guanine residue is cleaved) results in an abasic site
- DNA strand fragmented through  $\beta$  elimination, leaving 3'-aldehydic and 5'-phosphate ends

Ancient DNA:  
Few tens of base pairs



Meyer et al. 2017 eLife

DNA from fresh samples:  
Several thousands of base pairs



- Extensive fragmentation (typically < 100 bp)
- Ideal for high-throughput short-read sequencing (HTS)
- DNA extraction protocols suitable for ultra-short DNA fragments

## So, what is so special about ancient DNA?

- PROPERTY: Ancient DNA is heavily fragmented

EVIDENCE: Distribution of ancient DNA fragments length is skewed towards short fragments

- PROPERTY : ancient DNA is contaminated with exogenous DNA

EVIDENCE: Only a small fraction of the raw sequencing data aligns to the reference genome

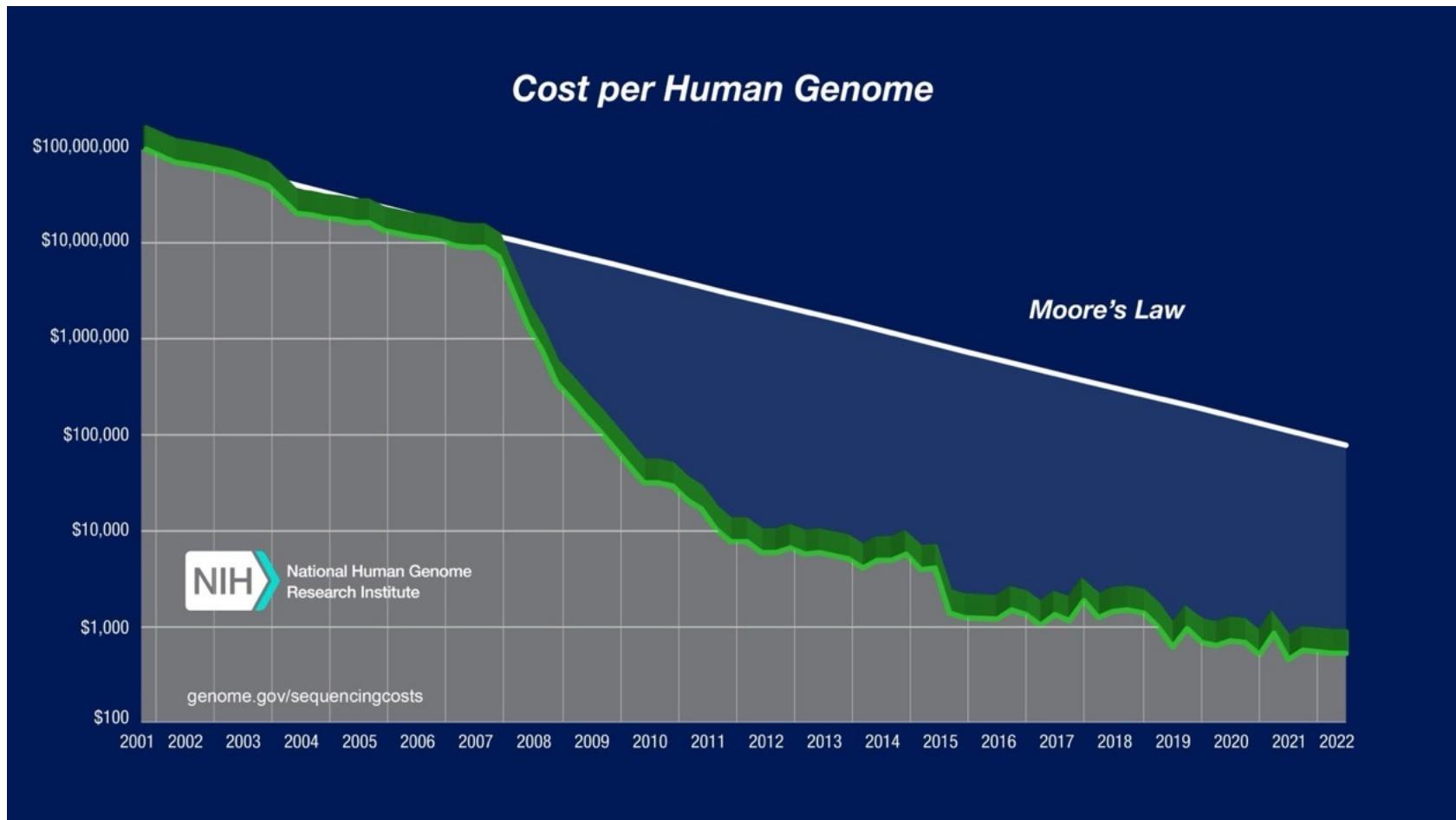
- PROPERTY : Cytosines are deaminated into uracils

EVIDENCE: Accumulation of C-to-T and complementary A-to-G substitutions at the ends of sequencing reads if DNA is not repaired with UDG and Endo VIII

- Introduction to ancient DNA
- Sources of ancient DNA and post-mortem DNA decay
- Properties of ancient DNA
- Ancient DNA analysis

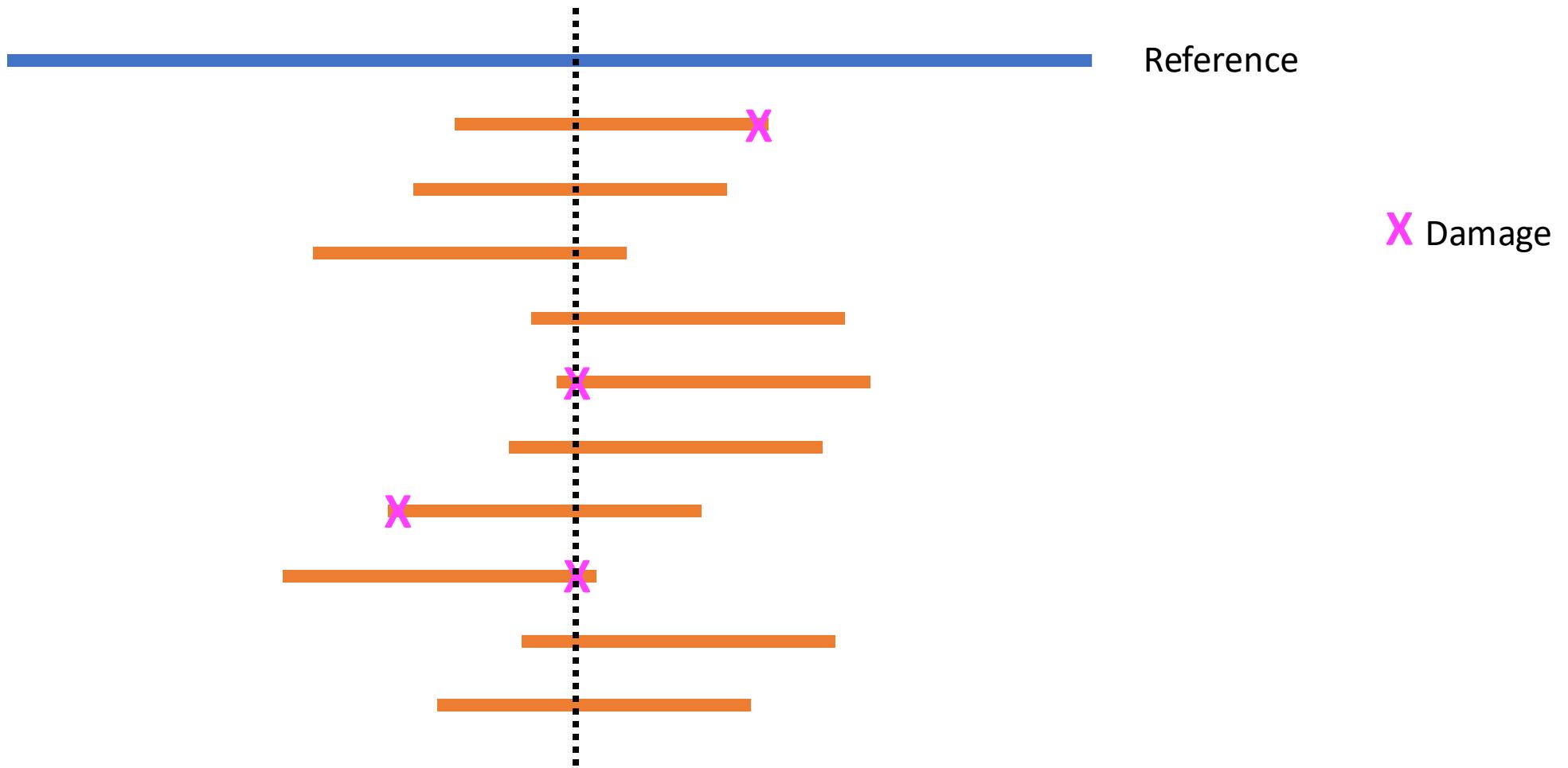
## High-throughput sequencing (HTS)

- Short-read sequencers generate DNA sequences  $\leq$  150 nucleotides
- Significant and sustained decrease in cost from mid-2000s



- HTS generate short DNA sequences
- Sequencing of virtually all molecules in an ancient DNA extract
- Independent replication for each genomic site

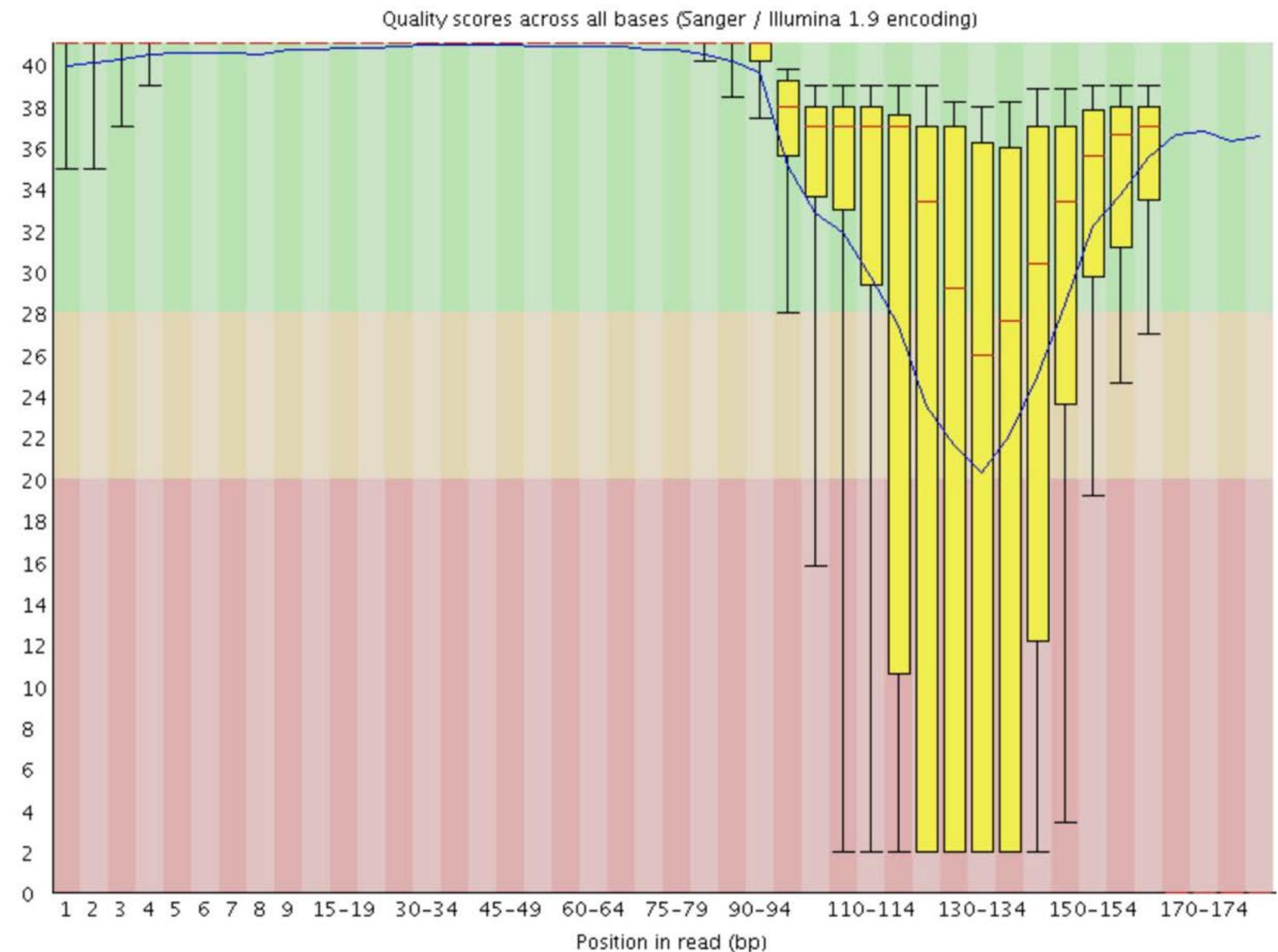
## Depth of coverage in HTS



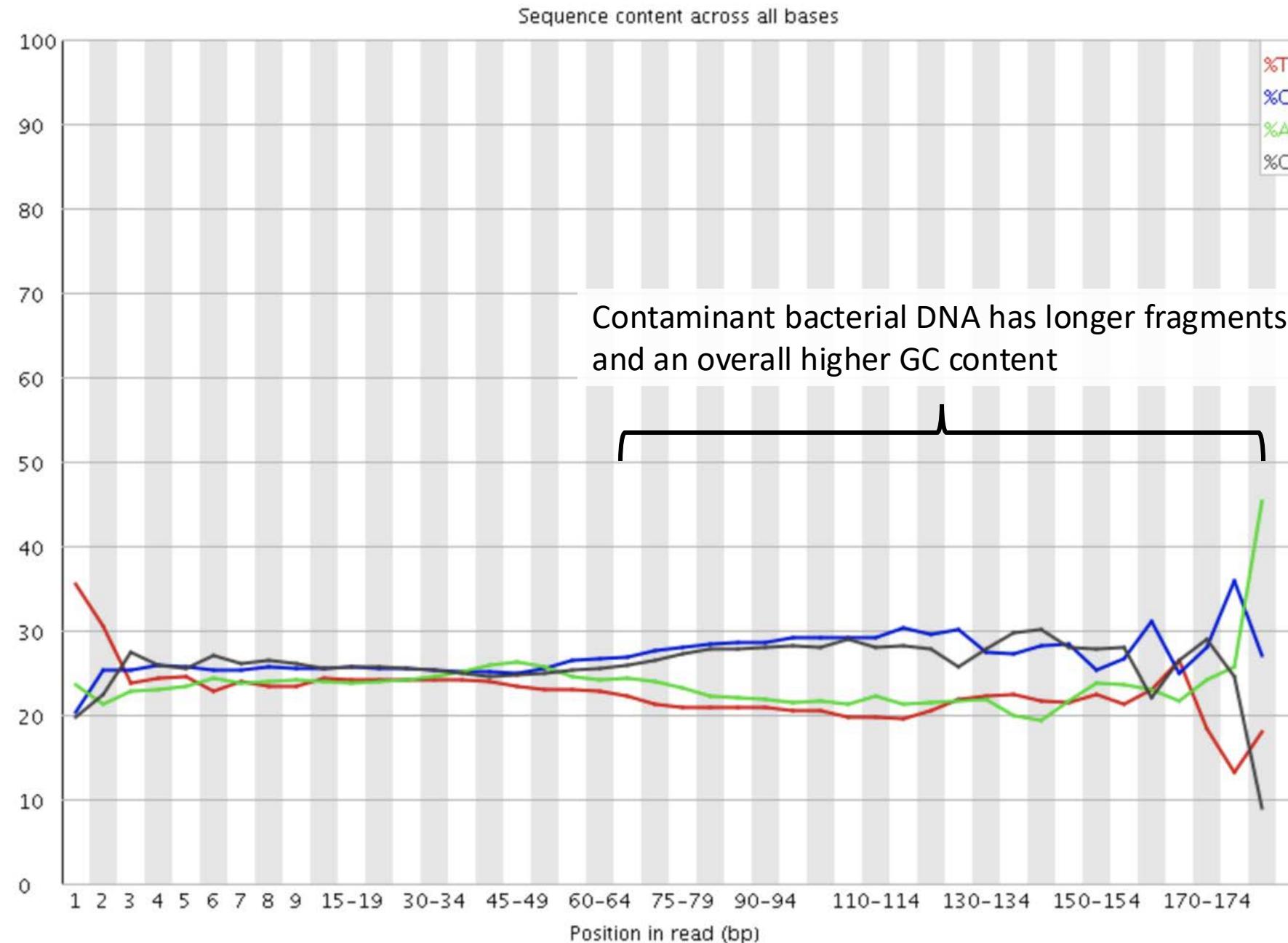
Deep sequencing of ancient DNA allows independent replication of variant sites to identify miscoding lesions and remove false positive results

## Typical base quality distribution of ancient DNA in Illumina sequencing data

- Sequencing runs are typically longer than ancient DNA fragments
- Many sequencing molecules fail to extend properly in the absence of template
- Errors accumulate and light signal is harder for the sequencer's camera to interpret correctly
- Base quality scores decrease dramatically towards the ends of ancient sequencing reads



## Typical higher GC content towards the end of sequencing reads

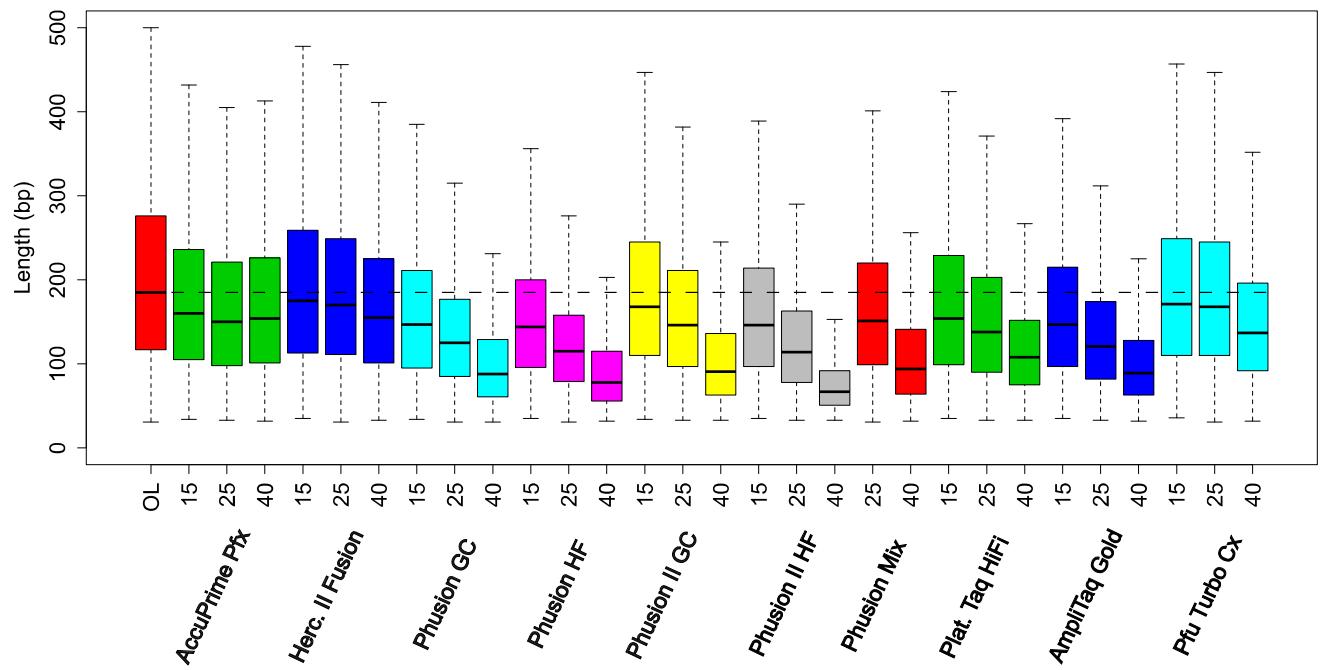
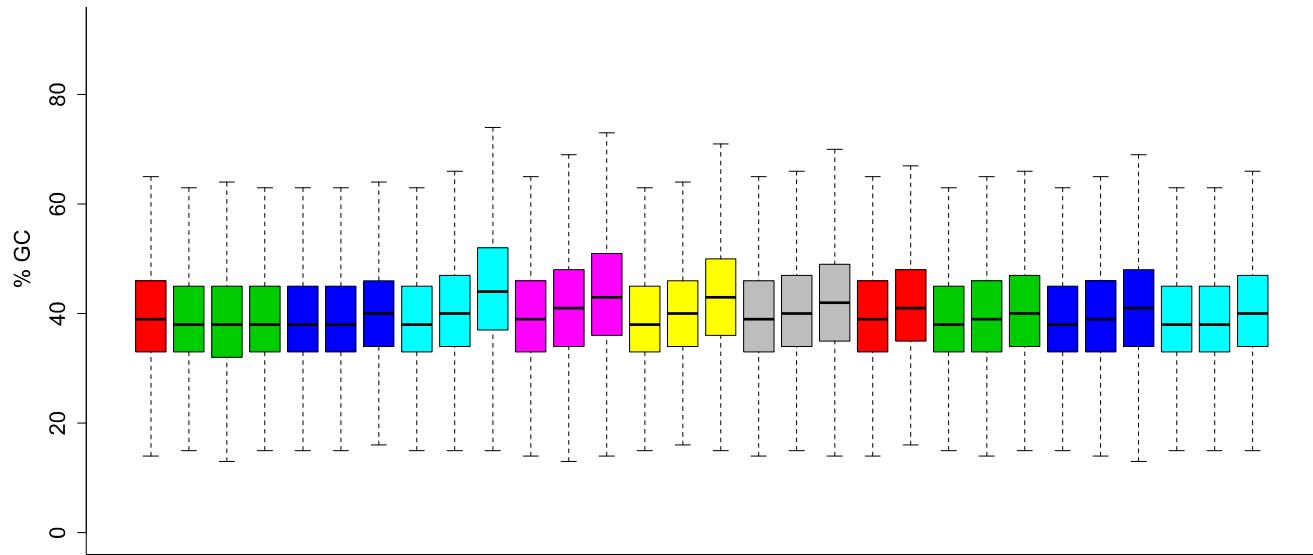


## Data biases due to experimental procedures

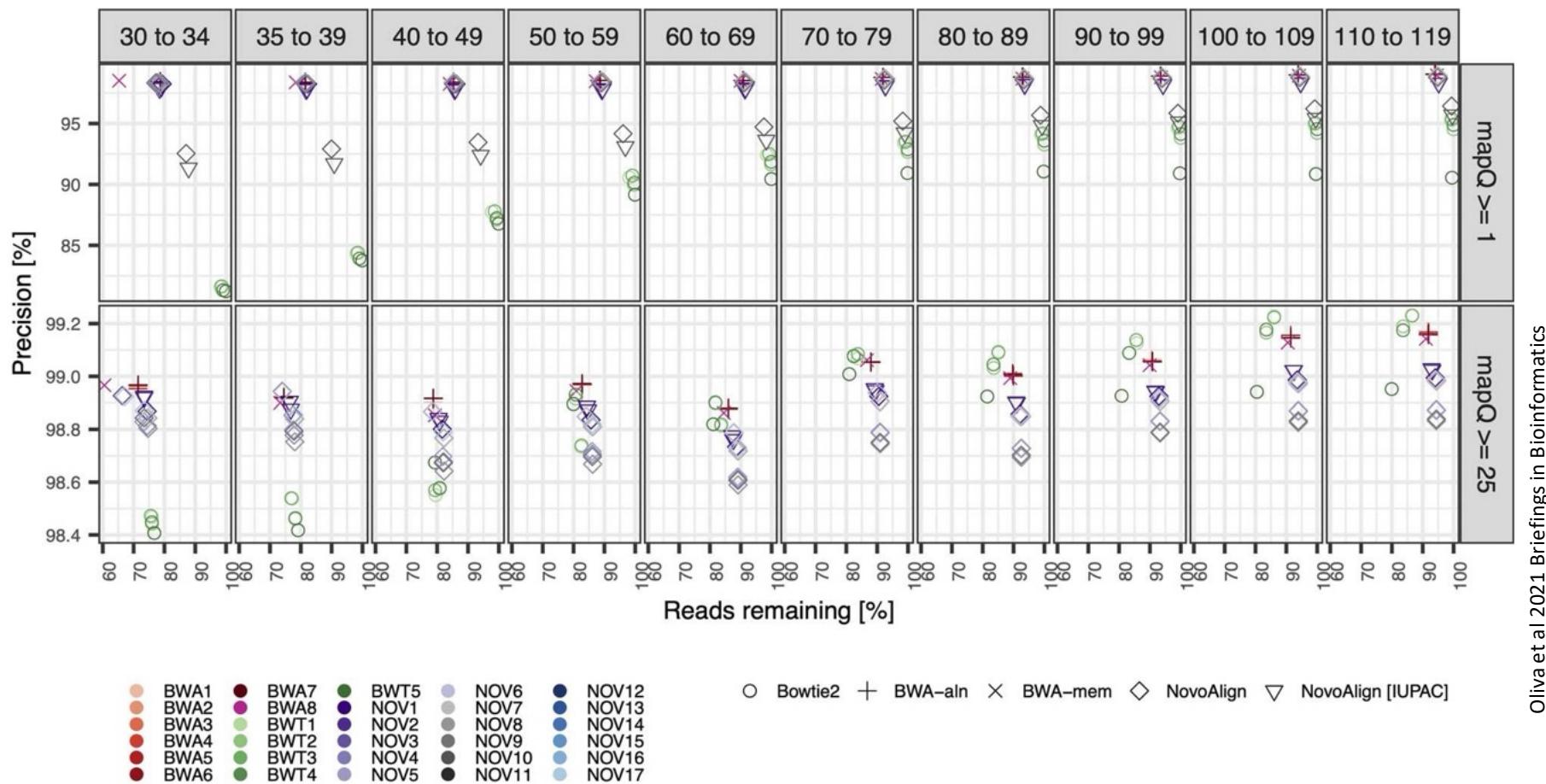
The performance of DNA polymerases used to perform PCR depends on GC content and fragment length

Some polymerases will amplify preferentially high GC-content DNA (i.e., microbial DNA)

Given the distribution of ancient DNA fragment lengths, length bias should be minimised



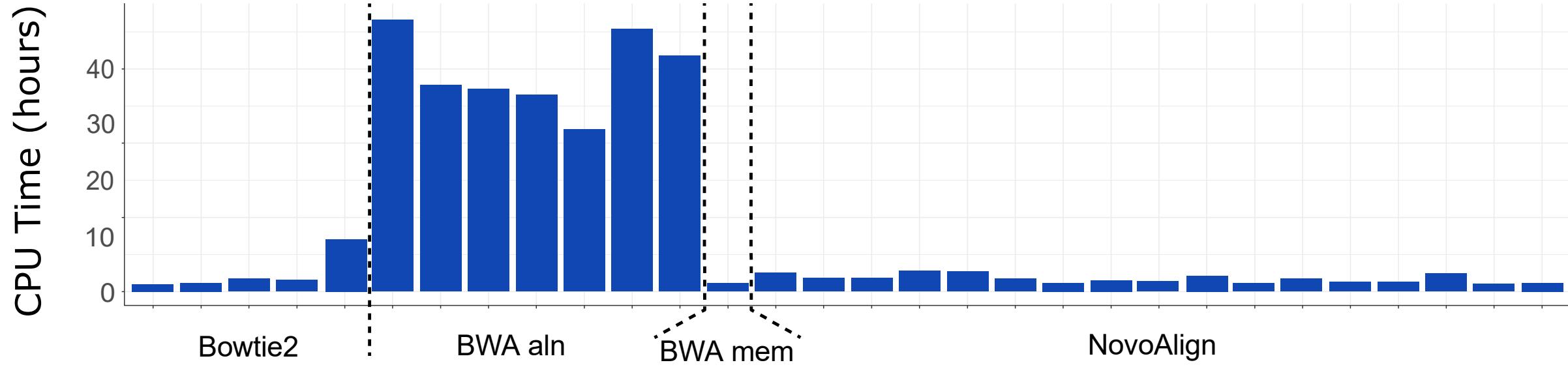
# Impact of read mapping strategies when processing ancient DNA sequencing data



- Due to the short length of ancient DNA reads, mapping software performance varies for:
- Precision, where the longer the read, the more accurate the mapping
  - The proportion of mapped reads

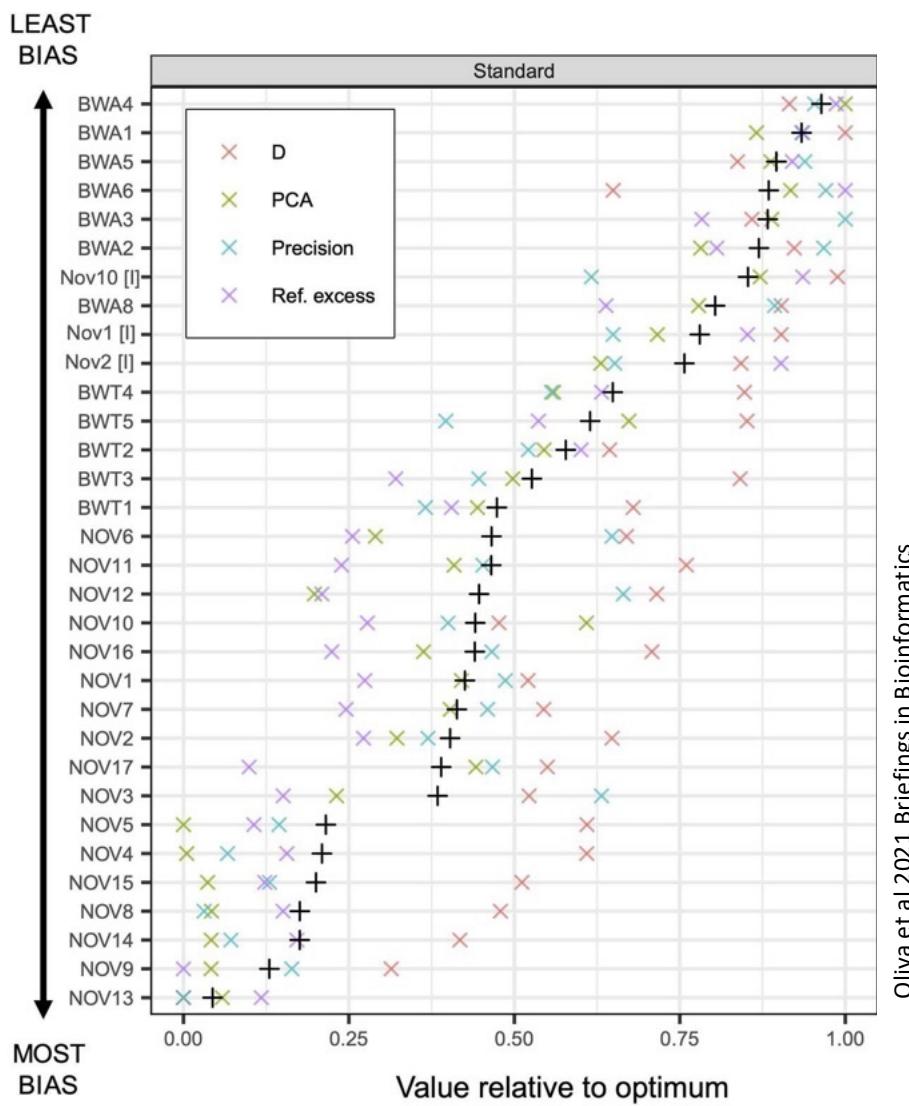
Filtering out reads with low mapping quality generally improves the mapping process

# Impact of read mapping strategies when processing ancient DNA sequencing data



Reads mapping strategy must take into account CPU time, especially if resources are limited

# Impact of read mapping strategies when processing ancient DNA sequencing data



- D: D-statistics estimate the level of relationship between test populations.
- PCA: PCA allows to visualise high complexity datasets in a 2D representation. The closer two individual datasets are in the PCA space, the more similar they are.
- Precision: mapping to the correct genomic location.
- Ref. excess: reference allele bias.

Reference bias leads to wrong population genetics inferences

## Summary

- Post-mortem processes impact DNA integrity and survival
- We can observe characteristic patterns in ancient DNA sequencing data
- It is possible to repair ancient DNA damage experimentally
- Bioinformatic methods can help estimate the amount of contamination
- Mapping of ancient DNA sequencing data requires balancing between key performance indicators



**Thank you for your attention!**

Don't forget to become a member of



<https://www.abacbs.org/>



<https://genetics.org.au/>



@DNATimeTravel



bastien.llamas@adelaide.edu.au